愛知淑徳大学論集－文学部篇－　第44号　2019.3　43-55

Linguistic contributions to the quantitative analysis of human language

若　山　真　幸
WAKAYAMA Masayuki

Key Words : sentiment analysis, semantic analysis, big data, AI, corpus study

1．Introduction

There are a number of opinions about whether modern linguistics is a science or not. Apparently, the answer is negative. It is generally said that scientific methodology includes 'observation', 'induction/generalization/prediction (hypothesis)', 'experiment for hypothesis', and 'analysis'. It is true that modern linguistic theories are quite abstract and based upon a number of assumptions which cannot be clearly verified. Consider language acquisition. It is quite amazing that children can master a language at a very early age without being taught by their parents or teachers. Within the principles and parameters framework (Chomsky (1986, 1991)), human beings are born with a finite set of principles that are common to all languages and a finite set of parameters that determine syntactic variation among languages like word order or the position of a *wh*-phrase. They strongly and logically claim, based on a number of observations and language data, that first-language acquisition is so fast due to a limited set of constraints for organizing language. It can be said that linguistics is a kind of science in this point.

Science begins with counting. Linguistics also begins with counting. Philologists have carefully studied the number of English expressions in written texts. Their extended efforts have so far explained how English was used in the past and how it has been changed over time. On the other hand, more and more researchers have introduced statistical approaches to linguistic analysis. One field is corpus linguistics, where a careful reading of concordance lines enables us to find linguistic patterns or grammatical rules that we have so far overlooked. Cooccurrence search is a good example. Another new trend is quantitative analyses of language. The use of big data has a big possibility to shed new light on not only language use and

also human behavior patterns.

In this paper, we will see how quantitative analyses of human languages have been developed and what they shed a light on. First, we will see the brief history of quantitative analyses of languages with a focus on corpus study. Then, it will be shown how big data have been used for sentiment analysis. Furthermore, based on our own corpus study, we will examine the relationship between basic emotions and their vocabulary. Finally, the present study will discuss how semantics and big data should cooperate with each other. In particular, this paper will argue that people do not use specific emotional expressions to express our emotions. Rather, it is essential to study collocational and semantic connections between five senses, polarity items, and sensory words in mining our opinions.

2. The brief history of quantitative analyses of languages

Word order in a sentence is not random. Syntax is one of the fields to describe and explain word order patterns in human language. According to the principles and parameters theory (Chomsky (1986, 1991)), word order in a given language is fixed by the directionality parameter (or the head-initial/final parameter) and the characteristics of the relevant features. On the other hand, corpus and computational linguistics also analyze a sequence of words in a given language. Unlike Chomskyan theories, which employs deduction in understanding our languages, the latter two fields quantitatively generalize language characteristics from a set of observations. In other words, the theories try to describe languages based on the frequency in use. In this section, we will briefly review two major recent trends of quantitative analysis of language.

2.1 Corpus linguistics

The term 'corpus' exclusively means a large set of written and spoken *electronic texts* (e-texts), which enable us to analyze language with a computer. One of the most famous linguistic corpora in the early days is the Brown Corpus[1], compiled in the 1960s. The collection of standard American English contains 500 samples across 15 genres, which were published in 1961, with approximately 1 million words[2]. In addition, the Helsinki Corpus of English is a diachronic corpus with about 1.5 million words. The corpus covers texts from c. 730 (Old English) to 1710 (Modern English). In this way, linguists were able to analyze languages both synchronically

[1] The original data was tape-recorded.

and diachronically. Obviously, computers can find a given expression more quickly and accurately than human eyes. This is one of the advantages of the corpus study. In order to analyze these text data, we are required to build concordance lines or the Key Word In Context (KWIC) with some analysis software. By generating concordance lines, we can easily find which word typically occurs with a target word in a frequent order. This makes it possible to explain a number of collocational patterns that we had not noticed. This is another merit of the corpus study.

As the IT technology advances, the size of corpus data has increased. For example, the BYU corpora[3] have released a variety of corpora, among which the Corpus of Contemporary American English (COCA)[4] contains 560 million words from 1990-2017. In addition, in May 2018, the organization released the iWeb corpus with 14 billion words in size. We can easily identify most common words in English with a number of mouse clicks, as shown in Table1.

RANK	WORD	FREQ
1	the	746,240,010
2	be	502,444,517
3	and	387,116,084
4	a	356,857,153
5	of	345,329,703
6	to	244,212,028
7	in	220,948,681
8	for	142,068,991
9	to	141,906,404
10	have	141,845,767

Table 1 ：10 most common words in English[5]

This is a well-known English fact, but just a list. In any case, the analysis of languages depends on reading concordance lines.

However, a new word prediction algorithm caught the eyes of computational linguists to see the frequency of co-occurrence of given words： an N-gram. We can sometimes predict future words in our utterance. This is because words or our languages are arranged in a probability of a sequence. Thus, calculating n-grams of a given language will reveal the likeliness of possible word or phrase sequences in

[2] The Lancaster-Oslo/Bergen (LOB) Corpus is a British English counterpart of the Brown Corpus with the same size of data.
[3] https：//corpus.byu.edu/corpora.asp
[4] the size of the corpus is based on 2018. (https：//corpus.byu.edu/coca/)
[5] iWeb corpus (https：//corpus.byu.edu/iweb/)

the language. Now, consider any sequence of N-words. As in the case of "*this is a nice sofa*", a 2-gram is a two-word sequence of words like "*this is*","*is a*","*a nice*", and "*nice sofa*", and a 3-gram is a three-word sequence of words like "*this is a*","*is a nice*","*a nice sofa*". These collected data provide a new English knowledge that we have not noticed. According to the Corpus of Contemporary American English (COCA)[6], the most 10 frequent bigrams in English are as follows.

	times of occurences		
1	2,551,888	of	the
2	1,887,475	in	the
3	1,041,011	to	the
4	861,798	on	the
5	676,658	and	the
6	648,408	to	be
7	578,806	for	the
8	561,171	at	the
9	498,217	in	a
10	479,627	do	n't

Table 2 ：the 10 most frequent bigrams in the COCA[7].

The most frequently used sequence is the preposition *of* followed by *the*. Table 2 clearly shows that the most frequently used word '*the*' is likely to occur with other functional words in its actual use. In addition, we can examine VERB + *the* + NOUN sequence in the same corpus. In our utterance, the definite determiner tends to occur with *open* and *tell*, and to be followed by a noun *door*.

RANK	3-grams	FREQ
1	opened the door	3,446
2	tell the truth	1,889
3	telling the truth	1,874
4	open the door	1,813
5	opens the door	1,471
6	closed the door	1,341
7	solve the problem	1,256
8	tell the story	1,238
9	change the way	1,168
10	use the word	1,108

Table 3 ：trigram of *the* (in the sequence of V+*the*+N) [8]

[6] https：//www.ngrams.info/download_coca.asp
[7] Davies, Mark. (2011) N-grams data from the Corpus of Contemporary American English (COCA). Downloaded from http：//www.ngrams.info on October 28, 2018
[8] https：//www.ngrams.info

Table 3 reflects more useful information of daily use of English than Table 1 and 2. As the length of n-grams becomes longer, therefore, we can obtain useful syntactic and semantic information on English expressions.

Corpora have a good advantage to see the tendency of language actual use. It is important to note that the corpus study has contributed greatly to the fields of applied linguistics such as language education, dictionary editions, too. So far, for example, dictionary editors subjectively determined the choice of words and their definitions based on their experiences and knowledge. On the other hand, most dictionaries are recently edited according to the frequency in use. Furthermore, corpora can clearly explain how native speakers use English on a daily basis and describe a number of differences between actual use of English and prescriptive grammar. Thanks to corpora, we can also figure out how and how many times Japanese learners of English make mistakes in their English compositions and speeches.

2.2 The roles played by Google

Google collects text data on a daily basis and has provided some tools based on the big data. Google Books Ngram Viewer[9] is very attractive for linguists in that it can visually demonstrate language change. The Ngram Viewer illustrates how many times a given expression occurs in a corpus of books over the selected periods. According to Michel and Liberman Aiden et al. (2011), the corpus is made of digitalized around 5 million books, with 361 billion English words (cf. Michel and Liberman Aiden et al (2011 : 176)). The books are collected from the 1500s to 2000s.

It is interesting to note that they mentioned two grammatical changes in their paper and talk[10]. First, they explained the change of the irregular verb form. For example, *thrive* was an irregular verb in that its past tense form did not follow the normal rule (e.g. affixation of *-ed*) but underwent a stem vowel change like *throve*. Like other many irregular verbs, *thrive* began to be a regular type at one point.

[9] https : //books.google.com/ngrams
[10] Lieberman Aiden and Michel (2014) What we learned from 5 million books (https : //www.ted.com/talks/what_we_learned_from_5_million_books)

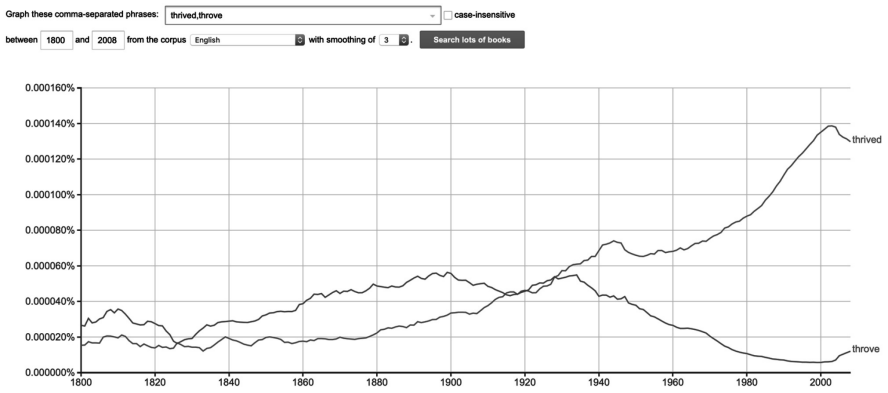

Figure 1 ：the change of the irregular form of *thrive*

As shown in Figure1, the number of *thrived* surpassed that of *throve* between 1920 and 1930. In addition, the irregular form is gradually decreasing still now. So far, we do not have a good approach to explain how and when irregular verbs turned into regular verbs. However, Google Books Ngram Viewer provides a clear answer to this question.

Their second example is the use of the *be*-verb for collective nouns. The noun '*the United States*' self-evidently means 'a country made of 50 states'. By the way, which form agrees with the noun : a singular or plural verb?

Figure 2 ：the use of the *be*-verb for the noun '*the United States*'

Actually, both forms are acceptable. Furthermore, preference is clear. Figure2 shows that '*the United States*' agrees with *it* rather than *are*[11] in the present-day

[11] They tried to relate the reason to the strong bond after the independence war against England. However, I leave it open here.

English.

There are some criticisms against the methodology of Google Books Ngram Viewer because of, for example, data bias of google books corpora and data accuracy caused by improper data-scanning. Nevertheless, this tool is quite helpful to see historical changes of English expressions.

2.3 Sentiment analysis[12]

Sentiment analysis examines people's opinions and emotions expressed in written texts. According to Liu (2015), sentiment is the underlying feeling, attitude, evaluation, or emotion associated with an opinion. There are main three sentiment orientations : positive, negative, and neutral. With respect to the last type, an opinion is taken to be neutral when it lacks any sentiment.

Sentiment analysis is associated closely with linguistics and psychology. Liu (2015, Chapter 2) argues that there are a number of problems with sentiment analysis. First, opinions (emotions) cannot be distinguished from one another. This is probably because language is ambiguous and polysemous. One clear example is that opinions holders can express their same opinions with the different expressions. Or, they might say what they feel differently with similar words. Second, there are different levels of sentiments. For example, *good* is weaker than *excellent*, and *dislike* is weaker than *detest*. English also has a set of intensifiers (e.g. *very*) and diminishers (e.g. *slightly*). It seems that they can express exactly how people feel. Obviously, it is almost impossible to evaluate the scale of our emotions in an accurate way. Indeed, sentiment analysis employs only three orientations, and recent star-based evaluation has five ratings (positive, fairly-positive, neutral, fairly-negative, negative). It can be said that we do not have to analyze our emotions in so much detail because language meanings are not clearly distinguished, or we recognize our emotions are not

Opinions are valuable information for various companies and industries in order to know how customer feel and think about their products and services. For example, we can see a great number of reviews at Amazon.com. There are positive, negative, and mixed reactions (or neutral opinions) toward almost all products ; people post a large number of reviews every day ; they sometimes give harsh words against products ; they also give very objective comments. By integrating application software, machine learning, statistics, and database systems, researchers

[12] See Liu (2015) for detailed information on the history of sentiment analysis.

can find out hidden patterns within written texts much faster than ever. This also makes it possible for companies to predict future, based on the current trends. Likewise, people can freely share their opinions about various topics in social media. Thus, public sentiments about topics on Twitter have drawn much attention from researchers. It is widely known that some researchers were able to predict the results of presidential elections in the United States only based on tweets by people[13]. Byrnes (2016) deals with researchers to claim that there were 'signals' on Twitter to show the trends.

3．How can semantics contribute to sentiment analysis?

We have seen that AI or big data is struggling to read our emotions. In this chapter, we will consider how semantics can contribute to sentiment analysis.

3.1 theory of emotions

How many emotions can we feel? To understand our emotions might be helpful to semantic analysis. It is widely believed that human emotions are universal and our innate ability. However, there have been many controversies over how many emotions we have. Like colors, human emotions are not clear-cut. Thus, it seems almost impossible to define the exact number of emotions. Along these lines, many researchers have discussed human emotions based on the proto-type theory. In other words, there are some basic emotions ; Other culture-specific emotions (if they exist) are derived from the basic types.

Darwin (1872) argued that we had a specific set of facial emotions based on the observation of facial expressions. Later, some psychologists revised his study and discussed six basic emotions.

(1) happy, sad, fear, anger, surprise, disgust

On the other hand, Plutchik (1980) proposed eight primary emotions and created a famous wheel of emotions, by which he described how emotions were related to each other.

(2) anger, fear, sadness, disgust, surprise, anticipation, trust, joy

[13] https : //www.technologyreview.com/s/603010/twitter-may-have-predicted-the-election/

Specifically, there are four basic emotions (joy, trust, fear, surprise) and four their opposites (sadness, disgust, anger, anticipation). They belong the primary emotions. In addition, there are secondary and tertiary emotions, and other combinations like '*hope*' (anticipation + trust).

A next question is what word we use in order to express such emotions. First, figure out the collocation of the verb '*feel*'. What adjectives does the verb occur with? I investigated the concordance line of the verb '*feel*' with given adjectives[14].

Rank	Words	Frequency	Polarity	Rank	Words	Frequency	Polarity
1	*good*	27.76 (914)	positive	16	*obliged*	11.14 (125)	
2	*sorry*	26.82 (730)	negative	17	*like*	11.00 (126)	positive
3	*guilty*	23.07 (540)	negative	18	*any*	10.87 (225)	
4	*comfortable*	20.41 (423)	positive	19	*threaten*	10.33 (110)	negative
5	*confident*	16.61 (284)	positive	20	*secure*	10.28 (108)	positive
6	*safe*	16.21 (275)	positive	21	*tired*	10.08 (106)	negative
7	*bad*	15.72 (283)	negative	22	*fine*	9.71 (111)	positive
8	*sick*	14.08 (205)	negative	23	*sure*	9.70 (126)	positive
9	*free*	13.94 (228)	positive	24	*terrible*	9.57 (98)	positive
10	*uncomfortable*	13.71 (190)	negative	25	*ashamed*	9.47 (91)	negative
11	*great*	12.82 (235)	positive	26	*ill*	9.38 (93)	negative
12	*right*	12.77 (205)	positive	27	*proud*	9.36 (95)	positive
13	*compel*	12.59 (159)		28	*my*	9.14 (258)	
14	*happy*	12.12 (172)	positive	29	*sad*	9.10 (89)	negative
15	*that*	11.75 (483)		30	*embarrassed*	8.95 (82)	negative

Table4 : the co-occurrence of the verb '*feel*' with an emotional adjective

It can be said from Table4 that we usually use *good, comfortable, confident, safe*, and so on, in positive meanings, while we tend to employ *sorry, guilty, bad, sick*, and so on, in negative meanings. Judging from emotional terms in (1) and (2), we usually do not use words in Table 2 to express our emotions. Next, look it differently. Let us examine a thesaurus to find synonymous words and expressions. (3) and (4) include the list of synonyms of *happy* and *sad*.

(3) Synonyms of *happy*[15]

cheerful (*703*), *contented* (*780*), *delighted* (*5109*), *ecstatic* (*429*), *elated* (*213*), *glad* (*3022*), *joyful* (*206*), *joyous* (*248*), *lively* (*1698*), *merry* (*694*), *overjoyed* (*251*), *peaceful* (*1953*), *pleasant* (*1982*), *pleased* (*4430*), *thrilled* (*1165*), *upbeat* (*805*)

[14] I made a collocation search on WordbanksOnline (provided by Shogakkan Corpus Network) and made a list of adjectival words which appear next to the verb feel. 'Frequency' in Table 2 means the number of occurrences per one million words.

[15] The list is collected from thesaurus.com (http : //www.thesaurus.com/) and the number of occurrences comes from WordbanksOnline.

(4) Synonyms of *sad*

bad (*36354*), *dark* (*9188*), *depressing* (*749*), *dismal* (*748*), *miserable* (*1376*), *moving* (*928*), *pathetic* (*818*), *pitiful* (*246*), *poignant* (*517*), *regrettable* (*182*), *serious* (*14373*), *sorry* (*8395*), *tragic* (*2374*), *unhappy* (*2368*)

Interestingly and surprisingly, words in Table4 do not appear as a synonym of *happy* in (3). In addition, only *bad* and *sorry* appear in both Table4 and Example (4). It is possible to assume from this that we do not use specific terms to express our emotions. Rather, we might use general polarity items, with broader meanings. Then, I collected synonyms of *good* and *bad*.

(5) Synonyms of *good*

acceptable, excellent, exceptional, favorable, great, marvelous, positive, satisfactory, satisfying, superb, valuable, wonderful

(6) Synonyms of *bad*

atrocious, awful, cheap, crummy, dreadful, lousy, poor, rough, sad, unacceptable

In Example (5) and (6), there are some words which are also used in Table 4. This result also shows that our choice of vocabulary is not so wide as the large number of synonyms of *good* and *bad*. This also means that we do not express our emotions precisely.

3.2 degree of likes and dislikes

Emotions are scalable and gradable. For example, Person A likes something, and Person B possibly likes it better than person A. On the other hand, Person C does not like something, and Person D hates it. In this case, the degree of dislike by Person B and D is clearly stronger than Person A and C, respectively. Then, how can we tell such a degree of emotions? There are at least three ways to do so.

(7) a. comparative and superlative
b. adverbs of degree, intensifiers, and diminishers
c. different vocabulary

First, comparative and superlative forms are straightforward. Second, adverbs of degree, intensifiers, and diminishers indicate the extent to which something happens. They are likely to modify extreme adjectives like *brilliant and awful*.

(8) *absolutely, completely, totally, utterly, really, exceptionally, quite, slightly, pretty, fairly, somewhat* …

Note that most of them are used exclusively in an affirmative or negative environment. Take (9) for example. In general, *Absolutely* and *exceptionally* modify adjectives with positive meanings like *right* and *brilliant*. On the other hand, completely and totally agree with adjectives with negative contents such as *different, wrong*, and *unacceptable*.

(9) a. positive polarity items : *absolutely, exceptionally, quite*
b. negative polarity items : *completely, totally, utterly*
c. both : *really, pretty*

Therefore, it is possible to tell or read the degree of emotions by using adverbs of degree and intensifiers. The most difficult way is to use different vocabulary as in (7c).

3.3 five senses and our sentiments

In the final section of Chapter 3, we will briefly mention the relationship between the five senses and our emotions. It is also possible to explore our emotions through the five senses because they are the initial point of how we feel. When we look something, we feel something. When we taste or smell something, we also react to the actions, which evoke some emotions inside us. Sensory words play an important role in describing five senses. In our corpus study, adjectives which tend to occur with basic perception verbs are collected to see how they are related to each other.

VERB		SENSORY WORDS
look	P	*good, great, comfortable, nice, stunning, cool, fantastic, bright, pretty*
	N	*tired, bad, dangerous, pale, bleak*
sound	P	*good, familiar, great, simple, wonderful, nice, interesting, right, easy*
	N	*strange, crazy, awful, ridiculous, odd, daft, silly, stupid, harsh, terrible*
touch		?
smell	P	*nice, good, sweet, fresh, lovely, wonderful, great, clean, delicious*
	N	*bad, awful, funny, sour, musty, horrible, unpleasant*
taste	P	*good, great, sweet, delicious, wonderful, nice, fine, amazing, fresh*
	N	*disgusting, horrible, awful, salty, terrible*

Table 5：co-occurrence relationship of five senses and emotional words[16]

It turns out, as in Table 5, that general polarity items like *good, great, nice, bad*, or *horrible*, appear more frequently than specific emotional terms like *bright* (sight) or *salty* (taste). As Winter (2016) points out, furthermore, smell and taste share a number of sensory words in both positive and negative meanings.

4．Final remarks

This paper has discussed the quantitative analyses of languages and some problems with sentiment analysis. Apart from data size and information technology, difficulties in sentiment analysis result from unclear boundaries between emotions, ambiguous characteristics of words, and their connections. It has been argued that semantic features can distinguish words from one another in semantics. Furthermore, such features have a big influence on grammar or syntax. Features are dichotomous, in that they are fully compatible with computational analyses in a binary world. Therefore, there is further discussion over the interactions between emotions and sensory words in the future.

References

Chomsky, Noam (1986) *Knowledge of Language : Its Nature, Origin, and Use*, New York : Praeger.

Chomsky, Noam (1991) "Some Notes on Economy of Derivation and Representation," in Robert Freidin, ed., *Principles and Parameters in Comparative Grammar*, 417-454, Cambridge, Mass. : MIT Press.

Darwin, Charles (1872) *The Expression of the Emotions in Man and Animals*, London : John Murray.

[16] The data is based on WordbanksOnline. I do not describe sensory words about '*touch*' because it is hard to define any specific verb for the sense of '*touch*'.

Liu, Bing (2015) *Sentiment Analysis : mining opinions, sentiments, and emotions*, New York : Cambridge University Press.

Michel, Jean-Baptiste et al. (2011) Quantitative Analysis of Culture Using Millions of Digitized Books, *Science* 331 (6014), pp. 176-182.

Plutchik, Robert (1980) *Emotion : Theory, research, and experience : Vol. 1. Theories of emotion*, *1*, New York : Academic Press.

Winter, Bodo (2016) Taste and smell words form an affectively loaded and emotionally flexible part of the English lexicon. *Language, Cognition and Neuroscience 31*, pp. 975-988.

Data and sources

Byrnes Nanette (2016)
https : //www.technologyreview.com/s/603010/twitter-may-have-predicted-the-election/ (MIT Technology Review)

the BYU corpora
https : //corpus.byu.edu/corpora.asp

Corpus of Contemporary American English
https : //corpus.byu.edu/coca/

Google Books Ngram Viewer
https : //books.google.com/ngrams

iWeb
https : //corpus.byu.edu/iweb/

Lieberman Aiden, Erez and Jean-Baptiste Michel (2014) *What we learned from 5 million books*
https : //www.ted.com/talks/what_we_learned_from_5_million_books

N-grams data from the Corpus of Contemporary American English (COCA)
https : //www.ngrams.info/download_coca.asp

WordbanksOnline (Shogakkan Corpus Network)
https : //scnweb.japanknowledge.com/WBO2/